基于深度森林的网络匿名流量检测方法研究与应用

李成豪1，魏松杰1

（1. 南京理工大学计算机科学与工程学院，江苏省南京市 210014）

**摘要：**流量分类一直是许多工作研究的主题，但互联网服务的快速发展和加密的普遍使用使其成为一个公开的挑战。加密对于保护互联网用户的隐私至关重要，这是近年来出现的各种隐私增强工具中使用的一项关键技术。其中，Tor匿名通信系统是最受欢迎的一种，基于匿名通信系统Tor构建的暗网是现今规模最大的暗网实体，常被犯罪分子用来从事各类违法活动，为犯罪分子提供了隐匿空间，给网络监管带来了十分严峻的挑战，因此高效识别Tor流量对网络监管和打击犯罪有着重要意义。本文根据Tor匿名通信流量特点设计了一组用于Tor流量行为检测的网络流特征，并在原有深度森林模型内存需求和时间开销大的不足基础之上设计了一种改进的深度森林模型用于Tor匿名网络流量的识别。实验结果表明，与已有识别方法相比，本文提出的模型准确率可达99.86%，并且节省了时间和内存。

**关键词 :** 流量分类 ; Tor ; 暗网 ; 匿名流量 ; 深度森林

**Research and application of Network Anonymous Traffic Detection Method based on Deep Forest**

LI Cheng-hao1, WEI Song-jie1

(1. *Department of Science and Engineering,* *Nanjing University of Science and Technology*, *Nanjing, Jiangsu Province* 210014, *China*)

**Abstract：**Traffic classification has been the subject of many studies, but the rapid development of Internet services and the widespread use of encryption make it an open challenge. Encryption is very important to protect the privacy of Internet users, which is a key technology used in various privacy enhancement tools in recent years. Among them, tor anonymous communication system is the most popular one. The dark network based on tor anonymous communication system is the largest dark network entity at present. It is often used by criminals to engage in various illegal activities, providing hiding space for criminals, and bringing severe challenges to network supervision. Therefore, efficient identification of tor traffic is of great significance to network supervision and combating crime. According to the characteristics of tor anonymous traffic, this paper designs a set of network flow characteristics for tor traffic behavior detection, and designs an improved deep forest model for tor anonymous network traffic identification based on the shortcomings of the original deep forest model, such as large memory requirements and time overhead. The experimental results show that, compared with the existing recognition methods, the accuracy of the proposed model can reach 99.86%, and save time and memory.

**Key words：**traffic classification; Tor; darknet ; anonymous traffic; deep forest

**0 引言**

随着互联网的飞速发展，互联网的规模不断扩大，逐渐渗透到社会、经济和政治的方方面面，而互联网中个人的安全和隐私保护问题日益严峻。匿名网络设计应运而生，其初衷虽然是为了保护用户的隐私，但往往被不法分子滥用，以逃避网络追踪，进而实施恶意网络活动。由于匿名通信网络使用加密、多节点转发和填充等多种流量混淆方法隐藏了原始流量的特点，这使得传统流量识别技术已经无法准确识别匿名流量。个人用户、企业和政府每天都会使用各种网络应用，每天都会产生数以亿计的网络流量，恶意流量很容易将自己隐藏在海量的流量中，进而发动攻击，给网络安全带来极大的威胁。因此研究如何高效准确的识别海量流量中具有恶意的网络行为，对于网络安全发展具有极其重要的科学研究意义。

Tor作为目前使用最广泛的匿名通信系统之一[1]，一直是学术界研究的主题之一。它在2004年推出了一项称为隐藏服务的技术，实现了客户端与服务端的双端匿名。通过使用隐藏服务技术，任何服务运营商都可以在提供网络服务的同时隐藏其服务器的物理地址。所以Tor拥有的用户最多，分布范围最广，平均每天的用户数量超350万人[2]。朴茨茅斯大学的Owen Gareth等研究人员在为期六个月的研究中观察到80%以上的隐藏服务的访问请求都指向已知的著名虐童网站。此外，枪支与毒品交易也是暗网市场不可或缺的部分。，但多数研究仅侧重于降低Tor的匿名性或提高其性能，对Tor流量的分析与识别的研究较少。因此对Tor匿名网络流量进行分析与识别具有很强的代表性。

由于Tor采用了TLS加密技术，检测者无法提取Tor流量内部的有效特征载荷，加密网络流量给网络安全防御带来了巨大的挑战，在不加解密的基础上识别Tor匿名流量具有十分重要的研究意义。Tor 虽然通过加密和封装等方式将非正常流量隐藏于普通流量，但是仍然可以基于Tor网络流特征的微小差异对其进行识别。随着机器学习的快速发展，越来越多的研究人员将机器学习技术应用于对Tor匿名流量的识别中，并取得了一定的研究成果，但也存在模型复杂度高、训练时间长、需要庞大的训练数据集作支撑、模型泛化能力差等方面缺陷。

通过阅读大量文献发现在已有的研究工作中，研究人员使用Netmate、pacp2flow、Tranalyzer等工具进行流量特征的提取，但这些工具多数都是基于数据明文字段进行计算，对于强加密的Tor匿名流量并没有很好的效果。通过研究Tor匿名通信系统本身的工作特性，发现以其通信频率、时延等为切入点计算时间相关特征具有很强的区分性，因此本文通过CIC Flow Meter处理Tor数据集产生的83中流量特征进行评估，最终选择重要性前35名特征作为Tor流量检测的时间相关特征[3]。为解决传统机器学习模型中存在的缺陷，本文通过构建改进深度森林模型最终取得了较高的精确率和召回率。

**1 相关工作**

迄今为止，国内外学者使用不同方法对网络流量分类进行了全方面的研究。早期通过Tor匿名通信系统的端口号、IP通信地址等进行识别。基于端口识别的分类方法在早期可以通过检查分组的传输层端口号[4]，再依据IANA（Internet Assigned Numbers Authority）定制的知名端口号与注册端口号列表将分组与应用对应起来。随着网络应用的不断更新发展Tor流量也对其数据包进行了加密处理无法检测其端口号，进而导致基于端口的Tor流量方法检测已不再适用，Moore等人通过实验发现，此方法的准确率不足70%。当前根据流量检测技术的原理和方法，将其分为基于深度包和机器学习的两个主要研究方向。

深度包检测技术（Deep Packet Inspection, DPI）最早在上个世纪的90年代被使用 [5]。这种方法在使用时不仅检测数据包的端口号，同时也对数据包的包头和包中的数据检测。使用深度包技术进行流量检测有两个主要的优点：首先是可以获得很高的准确率；其次在前期就检测出具有异常行为的网络流量。但于深度包检测技术有着计算复杂度高、无法应对加密数据、指纹获取困难、无法追踪到网络数据的位置等缺点。LIU[6]等人指出,利用深度包检测技术对加密流量进行检测,只有30%~70%的准确率。

人工智能的飞速发展，使得机器学习被应用到网络流量的分类与识别任务之中并取得了丰富的成果。机器学习方法应用到匿名流量检测系统之中，使其强大的数据学习和判别能力得到了充分的发挥，为网络特征流量检测增添了新的技术手段。主流的机器学习分类方法有支持向量机[7]、决策树[8]、贝叶斯方法[9]和随机森林[10]等。李晓明等人将贝叶斯网络算法、支持向量机算法和朴素贝叶斯算法在网络匿名流量识别上的精确率和召回率进行了对比[11]。机器学习方法应用到匿名流量检测系统的主要思想是：根据网络流量信息（网络服务类型、IP地址、协议类型、连接状态等）和主机信息（文件访问权限、主干资源访问、应用程序进程等）来建立相应的机器学习模型，进而对正常的流量和异常流量进行分类。

表征学习是深度学习的贡献之一，即通过端到端的训练，提取出数据集中更具有代表性的特征[12]。表征学习的主要作用是自动学习提取有利于识别的数据特征，对繁琐的原始数据简单化，将原始数据里面的无用信息去除，将有用信息更佳高效地进行提取。

综合以上研究现状，本文使用的深度森林[13]是一种以决策树为基础的集成学习方法，其在表征学习方面有着很强的能力。深度森林主要由级联森林和多粒度扫描两个组件构成。深度森林因为其级联森林组件，所以可以进行表征学习。在级联森林中，每一层都是由多个随机森林集成组成，每个森林都是由多个决策树集成组成，也就是说，级联森林的每层都是集成的集成。级联森林的每一层将上一层处理后得到的信息作为本层增广的输入信息。对于特征间存在关系信息的数据，深度森林使用另一组件多粒度扫描，使其表征学习的能力得到进一步的加强。深度森林不像深度学习，不需要复杂的超参数和庞大的数据集依然可以得到很高准确率。但是深度森林由于所有的样本都需要通过深度森林的所有层级，这就导致时间复杂度随层数线性变化。其次滑动窗口使每一个样本产生成百上千个新样本，使得计算代价大大增加。为了解决原始深度森林当前存在的问题，本文在原始深度森林的框架中增加了置信筛分机制[14]用来降低深度森林在时间和内存上的开销。

# 2 匿名流量检测系统

## 2.1 匿名流量检测流程

本实验首先将数据集进行特征提取和数据预处理等操作，然后通过改进的深度森林训练出模型，通过模型从混合流量中检测出隐藏在其中的Tor流量，如图1所示。

**图1 Tor流量检测流程**

## 2.2 Tor匿名流量特性分析

从Tor原始流文件中得到的数据集中包含非常多的空值、重复数据和高纬稀疏数据，将严重影响对Tor流量预测的准确性。本文通过分析Tor流量的特性，通过筛选用较少的特征便可以精准描述Tor流量行为。

用户通过Tor客户端发送请求信息，根据Tor匿名通信系统工作原理，用户的请求需要途经三个中间节点的转发才能被目标服务器接收，而且选取的三个节点通常部署在全球各地，因此Tor流量的转发时延比普通网络流量会大，可以通过数据包的时延以及流持续时间等特征描述转发时延差异。

通过研究发现Tor普遍将程序的数据包以固定的长度进行封装，所以Tor数据包长度是在一定范围内变化的。Tor流量和普通流量在数据包长度的特征上面存在显著差异，可以通过数据包长度的特征来观察数据包大小变化的情况及差异性。

此外，用户在建立Tor连接之后，如果在一段时间内应用程序不进行任何操作，网络变成空闲状态；如果应用程序开始工作，将不断的产生网络请求，网络将变成忙碌状态，产生的网络流的特征无法被Tor隐藏，可以通过流变成活动状态之前的空闲时间和流变成活动状态时的空闲时间的特征描述Tor这一特点。

综上所述， Tor 虽然通过加密和封装等方式将非正常流量隐藏于普通流量，但是Tor匿名网络的时延、流持续时间和数据包长度分布等特性仍然存在，通过对这些特征的处理，可以帮助区分Tor和NonTor流量。

## 2.3 Tor匿名流量特性分析

为了降低原始数据集中特征的冗余度，提高检测的准确率，本文在Tor和NonTor流量数据集上对CIC Flow Meter程序产生的83种流量特征进行了相关的重要性评估。

本文通过提取出一些流超时时间的数据集的基于时间的特征，将Tor数据集中的所有流标记为Tor，将其和相同流超时时间的加密NonTor数据集进行合并用作输入，并对每个流超时时间数据集（2s，4s，8s，16s，32s，64s，128s）应用多种不同的特征选择方法，最终Infogain+Ranker特征选择算法在处理16s数据集取得的特征实验结果最佳，如表1所示。

**表1 特征选择结果**

|  |  |
| --- | --- |
| 特征 | 相关度 |
| Flow Bytes/s | 0.8778 |
| Flow Duration | 0.8745 |
| Fwd IAT Mean | 0.7135 |
| Bwd IAT Mean | 0.7134 |
| Flow IAT Std | 0.7034 |
| Fwd IAT Min | 0.6961 |
| Flow IAT Mean | 0.6921 |
| Bwd IAT Min | 0.6874 |
| Idle Max | 0.6868 |
| Idle Min | 0.6755 |
| Idle Mean | 0.6726 |
| Flow Bytes/s | 0.6692 |
| **…** | **…** |

如上表所示，本文最终选择了特征相关度前35名的特征作为进行Tor匿名流量检测的特征。它们具有Tor匿名流量检测的最大信息增益和卡方统计量。数据集的特征属性及其描述如表2所示。

**表2 特征描述**

|  |  |
| --- | --- |
| 特征属性 | 描述 |
| Fwd IAT mean/max/min/std | 正向流中包到达间隔的 mean/max/min/std |
| Fwd Packet Length mean/max/min/std | 正向流中包长度的 mean/max/min/std |
| Bwd IAT mean/max/min/std | 反向流中包到达间隔的 mean/max/min/std |
| Bwd Packet Length mean/max/min/std | 反向流中包长度的 mean/max/min/std |
| Flow IAT mean/max/min/std | 包到达间隔时间的 mean/max/min/std |
| Active mean/max/min/std | 流变为空闲态之前活动时间的 mean/max/min/std |
| Idle mean/max/min/std | 流变为活动态之前空闲时间的 mean/max/min/std |
| Total Fwd Packet | 正向数据包总数 |
| Total Bwd packets | 反向数据包总数 |
| Total Length of Fwd Packet | 正向数据包总大小 |
| Total Length of Bwd Packet | 反向数据包总大小 |
| Flow Bytes/s | 每秒流字节数 |
| Flow Packets/s | 每秒流数据包数 |
| Flow Duration | 流持续时间 |

## 2.4 流量数据预处理方法

为了获得更加有效的Tor和NonTor流量数据集，本文需要对获得的数据集进行预处理。

### 2.4.1 消除重复数据

通过CIC Flow Meter设定流超时时间将不同类型的pcap文件保存到csv文件中并将其合并，通过研究发现无论是Tor数据集还是加密的NonTor数据集都存在大量的重复数据流，并发现有的重复率高达30%。这些重复数据不仅没有研究价值，反而容易使得实验结果虚高，因此本文首先对重复数据进行消除处理。

### 2.4.2 数据归一化处理

由于Tor数据集和NonTor数据集其中特征的取值范围跨度均较大，例如特征Flow IAI Max的取值范围为[1，15998544]，这将严重影响模型的收敛速度和分类性能。于是本文采用MIN\_MAX数据归一化方法，将所有特征范围归一化到[0，1]之内。

## 2.5 检测目标定义和效果评价方法

Tor流量行为检测是二分类实验，其主要目标就是从混合流量中检测Tor匿名流量。在匿名流量检测实验中，数据集中样本往往都是不平衡的。若正负样本比例差别非常大的情况下仍依靠准确率评判是不科学的。例如正常样本有99000条数据，异常样本只有1000条数据。在识别过程中即使所有的数据全判断为正常样本准确率也可以高达99%，此时准确率指标不仅不能表示性能很好，相反表示性能很差。本文使用准确率（Accuracy）、精确率（Precision）、召回率（Recall）和F1-Score等指标作为评价标准来综合判断Tor匿名流量检测的效果。

|  |  |
| --- | --- |
|  | （1） |
|  | （2） |
|  | （3） |
|  | （4） |

# 3 网络匿名流量检测模型设计

## 3.1 深度森林简介

深度森林是一种新颖的决策树集成学习[15]方法，具有很高的预测精度。深度森林主要由级联森林和多粒度扫描两个组件构成。深度森林因为其级联森林组件，其可以进行表征学习。每一层都是由多个随机森林集成组成，每个森林都是由多个决策树集成组成，即每层都是一个集成的集成。每层通过接受上层的结果后对其输入信息进行扩展。多粒度扫描通过使用滑动窗口对原始数据特征进行扫描，并通过不同尺寸的滑动窗口进行扫描，以获得更多不同的特征向量，进而增强级联森林的表征学习能力，使得深度森林模型对时序和空间特征数据的处理具有较好的效果。

## 3.2 改进深度森林模型设计

尽管Zhou等人[16]的通过实验证明通过增大深度森林的模型的复杂度将获得更好的效果，但随着多粒度扫描中使用更多不同尺寸的滑动窗口，以及级联森林的每层使用更多的决策树，同时深度森林模型的训练和预测需要的时间和内存的开销也就越大。导致此问题产生的有两个主要原因，首先所有的数据都要经过级联森林的每一层，使得时间开销随着级联森林的层数的增加线性增加；其次在每一个原始样本在经历多粒度扫描之后都会产生成百上千个新的样本，训练集被大幅度增加，使得计算开销剧增。

### 3.2.1 改进的级联森林

为解决所有样本都要经过级联森林的每一层而带来的时间和内存的开销，本文在级联森林结构引入了置信筛分机制[17]。置信筛分是将级联的每一层的实例分为两个子集：一个是容易预测的；另一个是难以预测的。如果一个实例很容易预测就直接当成最终结果输出，只有当一个实例很难预测时，它才需要被传递到下一层。改进级联森林如图2所示。

改进深度森林的级联结构如上图所示。基于置信机制的深度森林与原始深度森林主要的不同点有两个方面。首先改进的深度森林在级联的每一层都有门结构，将样本集划分为容易预测的样本和难以预测的样本两个子集。如图所示，具有高预测置信度的样本（Y）将直接输入到最后不需要传递到剩下的层级；只有预测置信度低于指定标准的样本（N）会被传递到下一层继续训练。其次，由于后续样本数量的不断减少，剩余样本的预测难度也越来越大。基于置信筛分机制的深度森林将逐层增加每个森林中决策树的数量用来增大森林模型复杂度来训练剩余的样本，而原始深度森林模型每层使用相同复杂度的深度森林模型。实验结果证明，随着层数的增加，从低到高增加级联森林的决策树数量可以使模型得到更好的泛化误差上界。

例如一个三分类问题，其中一个样本的预测类别的分布向量为，则把此样本的预测置信度设置为0.7。如图3所示。

**图2 改进的级联森林**

**图3 预测置信度生成示意图**

如何设置级联森林每一层的置信度阈值是一个关键的问题。本文使用的是阈值设置准则，使得级联森林的每一层可以自动确定自己的置信度阈值，使得深度森林模型在保证性能的同时，使其计算效率得到提升。在级联森林的第t层，其预测置信度阈值是根据第t层的交叉验证错误率来确定。令超参数表示高置信度的训练样本所需要达到的交叉验证错误率。根据预测置信度大小将当层训练样本按照降序排序，其中代表样本的预测置信度。置信度阈值的设置方式如下：

|  |  |
| --- | --- |
|  | （5） |

其中：

|  |  |
| --- | --- |
|  | （6） |

表示预测置信度最高的k个样本的交叉验证错误率。

### 3.2.2 改进的多粒度扫描

深度森林中的多粒度扫描过程会使得训练模型的内存和时间开销大大增加。本文采用采样多粒度扫描方法用来解决这个问题。在采样多粒度扫描中，首先使用滑动窗口生成新的变换的样本，再通过随机采样从新的样本集中获取一个子集用于接下来的深度森林模型的训练。本文只考虑二分类问题如图4所示，通过对新的数据集进行随机采样不仅使新样本的数量大幅度减少，同时把多粒度扫描过产生的特征的维度大幅度下降（从903维降低到90维）。随着级联结构层数的不断增加，难以预测的样本将被传到下一层，因此在级联结构每一层都使用采样多粒度扫描生成新的特征以增强其表示。

**图5 改进的多粒度扫描**

基于置信机制的深度森林主要由三个部分组成：置信筛分、可变模型复杂度和采样多粒度扫描。如当原始输入属性的特征维度是400，采样多粒度扫描中使用尺寸为100维的窗口用于滑动，设置采样规模大小为30。一个含有n个训练样本的数据集，使用此滑动窗口产生含有个新的训练样本的数据集，每个样本是100维的新特征。使训练集获得两种的森林（随机森林和完全随机森林）。在一个三分类任务，两个森林模型会产生180维的转换特征。生成的新特征被用于级联森林的输入，直至满足设置的终止条件后停止。

级联森林的每一层把混合流量数据集分割成两个子集。只有难以预测的样本集会被传入到下一层，并随着样本数量的逐渐减少，不断增大森林中决策树的数量。如果给定的测试样本的预测置信度大于当前层所设置的置信度阈值，则将对该样本的预测结果直接最为最终结果输出，否则继续将其传递到下一层进行预测。最坏情况下，此测试样本将通过级联森林的每一层在最后一层得到其预测结果。预测结果为预测类别的分布向量中的最大值。

# 4 实验

## 4.1 实验设置

为了验证本文算法的可行性，在表3所示的实验环境下进行实验。

**表3 实验环境**

|  |  |
| --- | --- |
| 类别 | 参数 |
| 处理器 | Intel(R) Core(TM) i5-6300 CPU |
| 显卡 | NVIDIA GeForce GTX 960M |
| Python3 | 3.7 |
| Tensorflow | 2.8.0 |
| Scikit-learn | 1.0.2 |

**参数设置**：在级联森林中每个层级包含4个随机森林和4个完全随机森林；每个森林通过3折交叉验证生成样本的类别分布向量。对于原始深度森林将每和森林设置100棵决策树。对改进深度森林第一层使用10棵决策树。随着层数的增加剩余训练样本数目不断减少，每个森林中的决策树会随着训练样本数目的不断减少而线性增加。若连续4层准确没有得到提升，则自动停止训练。

## 4.2 UNB-CIC Tor 数据集简介与处理

为了评估提出方法的有效性，本文使用了UNB（University of New Brunswick）发布的Tor流量有标签数据集来验证本文所提出的方法。科研人员首先在计算机中组成两个虚拟机：网关和工作站，如图6所示。由于所有Tor流量必须通过网关才能到达工作站，因此Tor流量变得透明。在网关处将Tor数据包进行捕获，Non-Tor流量在工作站处被捕获。科研人员通过利用Wireshark和tcpdump等工具进行捕获流量，总共生成了22 GB的PCAP文件。

**图6 Tor流量捕获场景**

本文中使用ISCXFlowMeter应用程序进行pcap数据包解析。按照具有相同目的 IP，目的端口，源 IP，源端口，协议> 值的数据包定义为流。FlowMeter 生成双向流，通过第一个数据包确定正向（从源到目的地）和反向（从目的地到源），因此统计时间相关的特征也在正向和反向分别计算。注意，TCP 流通常在连接断开时终止（由 FIN 数据包），而 UDP 流则由流超时终止。流超时值可以由单独的方案任意分配。

**表4 Tor流量检测数据分布**

|  |  |  |  |
| --- | --- | --- | --- |
| 流超时值 | Tor | nonTor | 总计 |
| 2s | 32901 | 162552 | 195453 |
| 4s | 17556 | 137155 | 154711 |
| 8s | 9713 | 102970 | 112683 |
| 16s | 5329 | 86495 | 91824 |
| 32s | 2924 | 69633 | 72557 |
| 64s | 1626 | 61012 | 62638 |
| 128 | 930 | 57655 | 58585 |

本文目标是从混合流量中检测出Tor流量，因此将数据集划分为Tor流量和非Tor流量。在本文中研究了几个流超时值，以确定流超时对最终结果的影响。特别是，我们将流的持续时间设置为2、4、8、32，64和128秒。表4显示了在不同流超时值下，对Tor流量行为检测提取的数据分布。

## 4.3 实验结果

本节实验使用流超时时间为2、4、8、32，64和128秒的经过预处理的数据集作为输入，按照2: 8的比例划分了训练集和测试集，使用了准确率、精确率、F1-Score、召回率、训练时长、测试时长和内存开销等标准作为模型性能的评价指标。使用原始深度森林模型gc和改进深度森林模型gcCS的Tor流量检测对比实验结果如表5所示。

**表5 不同流超时值下面的Tor流量检测实验**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 方法 | 准确率% | 精确率% | F1-Score% | 召回率% | 训练时长s | 测试时长s | 内存开销M | 流超时时间 |
| gcCS | 99.90 | 99.24 | 98.96 | 98.68 | 189.54 | 11.94 | 24.67 | 2s |
| gc | 99.90 | 99.62 | 99.20 | 98.78 | 692.11 | 29.06 | 25.14 | 2s |
| gcCS | 99.85 | 99.56 | 99.40 | 99.24 | 197.87 | 3.89 | 20.84 | 4s |
| gc | 99.87 | 99.67 | 99.48 | 99.3 | 467.00 | 21.25 | 25.55 | 4s |
| gcCS | 99.83 | 99.62 | 99.46 | 99.54 | 159.12 | 3.20 | 20.02 | 8s |
| gc | 99.85 | 99.68 | 99.59 | 99.51 | 378.96 | 20.94 | 23.125 | 8s |
| gcCS | 99.86 | 99.82 | 99.72 | 99.62 | 114.92 | 2.72 | 20.00 | 16s |
| gc | 99.87 | 99.82 | 99.68 | 99.75 | 291.86 | 32.57 | 22.60 | 16s |
| gcCS | 99.85 | 98.61 | 97.61 | 96.93 | 213.00 | 8.76 | 20.14 | 32s |
| gc | 99.89 | 98.96 | 98.37 | 97.78 | 259.00 | 29.43 | 23.09 | 32s |
| gcCS | 99.85 | 98.76 | 96.67 | 94.66 | 162.98 | 6.42 | 20.09 | 64s |
| gc | 99.85 | 98.46 | 96.68 | 94.96 | 148.86 | 17.75 | 21.13 | 64s |
| gcCS | 99.85 | 94.39 | 95.36 | 96.35 | 166.86 | 2.80 | 22.04 | 128s |
| gc | 99.90 | 96.88 | 96.88 | 96.88 | 141.00 | 24.55 | 22.18 | 128s |

由表5的实验结果可知随着流超时时间越来越大数据集越不平衡（数据集中Tor流量越来越少）导致精确率、F1-Score和召回率均出现了下降。原始深度森林和改进深度森林模型在预测效果不相上下，但改进后的模型花费的训练时长、测试时长和内存更少，因此改进深度森林较原始深度森林模型用更少的计算资源得到了更好的预测性能。

## 4.4 对比实验

为了验证改进深度森林在Tor流量检测的优势，通过综合考虑选择流超时时间为16s的数据集和多种常见机器学习模型进行对比实验。本文选择了决策树、随机森林、支持向量机和深度神经网络四种机器学习方法进行对比。实验结果如表6所示。

**表6 对比实验结果**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 分类器 | 准确率% | 精确率% | F1-Score% | 召回率% |
| gcCS | 99.86 | 99.82 | 99.72 | 99.62 |
| gc | 99.87 | 99.82 | 99.68 | 99.75 |
| SVM | 98.43 | 97.00 | 94.66 | 92.43 |
| RandomForest | 99.79 | 99.64 | 99.31 | 98.98 |
| DecisionTree | 99.69 | 99.07 | 98.99 | 98.9 |
| DNN | 98.49 | 97.47 | 94.85 | 92.37 |

由表6可知,原始的深度森林模型和改进的深度森林模型取得的分类效果都略高其他方法。综合表5和表6可知，改进的深度森林模型较原始深度森林无论在准确率和时间空间开销方面还是和其他方法的测试精度对比中都取得了良好的效果。

# 5 总结与展望

本文主要对Tor匿名通信系统产生的网络流量进行检测与识别。首先，通过分析Tor匿名流量与正常普通流量的特征，通过特征选择算法，本文提出了一组基于时间特征来检测并描述Tor流量的数据集。Tor匿名通讯系统虽然使用了多种混淆加密技术，但是其数据的数据包到达间隔时间、流持续时间和流变为空闲态之前活动时间等时间特征与正常流量存在较大差异。本文对UNB-CIC Tor 数据进行Tor网流量特征提取与流量检测实验，并使用改进的深度森林进行训练与学习。当流超时值为16s时实验结果最佳，在测试集中Tor检测的准确率高达99.86%，改进的深度森林较原始深度森林可以利用较少的计算资源获得较高的预测性能，即改进后的深度森林模型预测准确率更高。本文提出的Tor匿名流量检测系统虽然可以准确的检测Tor流量，但是在线检测功能尚未实现，未来将在该方面进行研究与改进。

# 参考文献

[1] Dingledine R, Mathewson N, Syverson P. Tor: The second-generation onion router[R]. Naval Research Lab Washington DC, 2004, 46(4): 337-337.

[2] Saputra F A, Nadhori I U, Barry B F. Detecting and blocking onion router traffic using deep packet inspection[C]//2016 International Electronics Symposium (IES). IEEE, 2016: 283-288.

[3] Lashkari A H, Draper-Gil G, Mamun M S I, et al. Characterization of tor traffic using time based features[C]//ICISSp. 2017: 253-262.

[4] Madhukar A, Williamson C. A longitudinal study of P2P traffic classification[C]//14th IEEE international symposium on modeling, analysis, and simulation. IEEE, 2006: 179-188.

[5] Moore A W, Papagiannaki K. Toward the accurate identification of network applications[C]//International workshop on passive and active network measurement. Springer, Berlin, Heidelberg, 2005: 41-54.

[6] Liu X B, Yang J H, Xie G G, et al. Automated mining of packet signatures for traffic identification at application layer with apriori algorithm[J]. J Commun, 2009, 29(12): 51-59.

[7] Este A, Gringoli F, Salgarelli L. Support vector machines for TCP traffic classification[J]. Computer Networks, 2009, 53(14): 2476-2490.

[8] Sahu S, Mehtre B M. Network intrusion detection system using J48 Decision Tree[C]//2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI). IEEE, 2015: 2023-2026.

[9] Helman P, Veroff R, Atlas S R, et al. A Bayesian network classification methodology for gene expression data[J]. Journal of computational biology, 2004, 11(4): 581-615.

[10] Dogru N, Subasi A. Traffic accident detection using random forest classifier[C]//2018 15th learning and technology conference (L&T). IEEE, 2018: 40-45.

[11] LI Xiaoming, Ren Hui, YAN Jinyao . Analysis and research on network traffic classification algorithm based on machine learning [J] Journal of Communication University of China: Natural Science Edition, 2017, 24(2): 9-14.

李晓明, 任慧, 颜金尧. 基于机器学习的网络流量分类算法分析研究[J]. 中国传媒大学学报: 自然科学版, 2017, 24(2): 9-14.

[12] Wang P, Ye F, Chen X, et al. Datanet: Deep learning based encrypted network traffic classification in sdn home gateway[J]. IEEE Access, 2018, 6: 55380-55391.

[13] Zhou Z H, Feng J. Deep forest[J]. arXiv preprint arXiv:1702.08835, 2017.

[14] Pang M, Ting K M, Zhao P, et al. Improving deep forest by confidence screening[C]//2018 IEEE International Conference on Data Mining (ICDM). IEEE, 2018: 1194-1199.

[15] Blumberg B, Downie M, Ivanov Y, et al. Integrated learning for interactive synthetic characters[C]//Proceedings of the 29th annual conference on Computer graphics and interactive techniques. 2002: 417-426.

[16] Zhou Z H , Feng J . Deep Forest: Towards An Alternative to Deep Neural Networks[J]. 2017.

[17] Ma P, Wu Y, Li Y, et al. DBC-Forest: Deep forest with binning confidence screening[J]. Neurocomputing, 2022.